

Memory biased random walk approach to synthetic clickstream generation

Nino Antulov-Fantulin
Division of Electronics, Ruđer
Bošković Institute, Zagreb,
Croatia
nino.antulov-
fantulin@irb.hr

Matko Bošnjak
Department of Informatics
Engineering, Faculty of
Engineering, University of
Porto, Portugal
Division of Electronics, Ruđer
Bošković Institute, Zagreb,
Croatia
matko.bosnjak@gmail.com

Vinko Zlatić
Theoretical Physics Division,
Ruđer Bošković Institute,
Zagreb, Croatia
ISC-CNR, Dipartimento di
Fisica, Università "La
Sapienza", Roma, Italy
vinko.zlatic@irb.hr

Miha Grčar
Department of Knowledge
Technologies / E8, Jožef
Stefan Institute, Ljubljana,
Slovenia
miha.grcar@ijs.si

Tomislav Šmuc
Division of Electronics, Ruđer
Bošković Institute, Zagreb,
Croatia
tomislav.smuc@irb.hr

ABSTRACT

Personalized recommender systems rely on personal usage data of each user in the system. However, privacy policies protecting users' rights prevent this data of being publicly available to a wider researcher audience. In this work, we propose a memory biased random walk model (MBRW) based on real clickstream graphs, as a generator of synthetic clickstreams that conform to statistical properties of the real clickstream data, while, at the same time, adhering to the privacy protection policies. We show that synthetic clickstreams can be used to learn recommender system models which achieve high recommender performance on real data and at the same time assuring that strong de-minimization guarantees are provided.

Categories and Subject Descriptors

K.4.1 [Computers and society]: Public Policy Issues—*Privacy*; H.2.8 [Database management]: Database applications—*Data mining*; G.3 [Mathematics of Computing]: Probability and statistics—*Markov processes*

Keywords

memory biased random walks,
information networks,
recommender systems,
synthetic clickstream generation,
clickstream anonymization

1. INTRODUCTION

Advances in information technology constantly increase the quantities of information produced, transmitted and stored in each second around the world. This drives researchers to constantly improve the ability and performance of various data mining algorithms. Modern data mining algorithms are also able to easily extract unique patterns from data and thus pose serious threat for breaching privacy protection policies.

Recommender systems proved to be a very useful personal decision support tool in search through vast amounts of information on the subject of interest. Recommender techniques [1], [2] are divided into three main categories: content-based, collaborative-based and hybrid-based recommender techniques. Content-based techniques [3], [4] recommend items with respect to available structured or unstructured knowledge about user's online history; collaborative filtering techniques [5] recommend items with respect to past behaviour and interaction of all of the users and all of the items, whereas hybrid-based techniques [6] use combined approaches of the previous two techniques.

Regardless of the used technique, the system must have access to the users' private data - data consisting of users' history on the website and possibly other personal information. The quality of data, meaning the level of details, dictates the efficacy of the system. In the first approximation clickstream is a sequence of clicks made by a particular user while web browsing, leading from one item of interest to another. Clickstream data are considered to be a personal information, therefore privacy policies heavily restrict the public availability of this data, often rendering it unavailable for research.

To overcome these privacy policies, many real data sets (Netflix dataset, AOL search logs, Massachusetts Group Insurance Commission (GIC) medical encounter dataset, etc.) were anonymized by removing all the explicit personal identification attributes like name, social security number, etc.

Nevertheless, successful real world 'linkage attacks', on these data sets have been made, where researchers manage to identify personal records by linking different datasets by quasi personal identifiers like search logs, movie ratings, gender, ZIP code, etc. In case of the Netflix challenge dataset all explicit personal identification attributes were replaced with unique random number, but still the researchers [7] managed to link the Netflix dataset with the IMDB dataset by dates of user's ratings and partly de-anonymize the Netflix dataset. Dataset published by Massachusetts Group Insurance Commission (GIC) was also a target of linked attack by researchers [8] who exploited the fact that it is not common for two persons to have the same zip code, gender and birth date. Sweeney showed [8] that 87% of the U.S. population are unique with respect to the quasi identifiers like ZIP code, gender and date of birth. In August 2006, AOL published dataset of around 20 million search queries of 650 000 users from three months period which contained users clickstreams (URLs) from search results. Only a few days later, the New York Times reporters linked one clickstream with a real user which triggered a lawsuit against the AOL in the U.S. District Court in September 2006.

The paper is organized as follows. In Section 2 we present related work on problem of privacy-preserving data publishing. In Section 3 we describe data generator matrices used for the generation of synthetic clickstreams and the biased and unbiased random walk models on graphs. In Sections 4 and 5 we follow up the previous section and apply a biased random walk model to clickstream data. We focus on memory biases and user-sampling procedure to construct clickstream synthetic datasets. In Section 6 we present experiments performed to show applicability of our method in recommender systems. We provide additional anonymization guarantee for synthetic clickstream datasets in Section 7.

2. RELATED WORK

The problem of privacy-preserving data publishing [27] [28] and privacy-preserving data mining [29] are intensively researched within three research communities: database community, statistical disclosure community and cryptography community. Different privacy protection models have been proposed in order to counter possible privacy attacks like record linkage, attribute linkage, table linkage and probabilistic attack.

Record linkage models like k -Anonymity model [30] [31] [32] assure that the number of records with some quasi-identifier id is at least k and therefore assure the value of linkage probability at most $1/k$. Attribute linkage models like L -diversity [33] [34] are envisioned to overcome the problem of inferring the sensitive values from k anonymity groups by decreasing the correlations between quasi-identifiers and sensitive values. L -diversity model assures that the entropy of sensitive attributes in each group is larger than some threshold value l . A high value of l implies smaller probability of inferring sensitive values. Probabilistic models like ϵ -differential privacy model [35] ensure that the difference between the prior and posterior beliefs is small enough. The ϵ -differential privacy model ensures that individual's presence or absence in the database does not effect the query output significantly. Post-random perturbation (PRAM) methods [36] [37] change original values through probabilistic mechanisms and thus, by introducing uncertainty into data, reduce

the risk of re-identification.

Synthetic data generation is an alternative approach to data protection in which the model generates synthetic dataset preserving the statistical properties of original dataset. Several approaches for synthetic data generation have been proposed: (i) synthetic data generation by multiple imputation method [38], (ii) synthetic data by bootstrap method [39] (estimating multi-variate cumulative probability distribution, derive similar c.d.f. and sample synthetic dataset), (iii) synthetic data by Latin Hypercube Sampling [40] (multivariate synthetic dataset), (iv) and others such as a combination of partially synthetic attributes and real non confidential attributes [41] [42].

Most of the aforementioned anonymization strategies were developed for database records with fixed number of attributes but not for sequences such as clickstream data. This led us to propose a method for synthetic clickstream generation based on random walks. Random walks [9], [10], [11], [12] have been previously used for construction of recommender systems for different types of graph structures emanating from users' private data, but not for generation of synthetic clickstreams. We propose an approach for synthetic clickstream generation by constructing a memory biased random walk model (MBRW) on the graph of the clickstream sequences, which is a subclass of Markov chains [13], [14]. We also use the MBRW model to generate synthetic clickstreams with similar statistical properties to the real clickstreams. The MBRW algorithm can be understood as a clickstream generation and also an anonymization process. Re-identification is harder and uncertain due to the fact that the synthetic clickstreams are results of discrete stochastic process. Furthermore, we can provide even stronger privacy protection guarantee w.r.t. (Θ, ω) de-anonymization definition [7].

Many data-privacy researchers state that high dimensional data poorly resist to de-anonymization [7] which is a serious problems for companies, and prevents usage of real-life datasets for research and for data-mining challenges. Currently, 1 million dollar worth Overstock.com recommender challenge [44] is running in which synthetic data, which shares certain statistical properties with real data sets, was released. They state that instead of releasing sensitive data, they can bring the recommender code to the data in the cloud. In the final round of the challenge contestants codes will be uploaded to RecLabs [45] core server to build models and to evaluate them against the real data. They only claim that their synthetic data may share certain real statistical properties and should be used just to test if code works. It would be useful both for contestants and the company if synthetic data could be also used as a precursor for model performance. In the rest of the paper we demonstrate that synthetic data generated with our method are indeed a good testing ground for recommender systems.

3. RANDOM WALK MODELS

Data generator matrices

Clickstream is a sequence of web pages or items in general visited by a user. A clickstream c^i is an ordered sequence of web pages $c^i = u_1^i, u_2^i, u_3^i, \dots, u_k^i$, visited by a particular user i . If the web page u_j^i was visited before the web page u_k^i by user i , then $j <_i k$. The inequality relation "less than": $<_i$ is a function of particular user i . The set of all the clickstreams

in a system is $C = \{c^1, c^2, \dots, c^i, \dots, c^n\}$.

We define two characteristic matrices for the set of clickstreams C : (i) Direct Sequence Matrix (DS) and (ii) Common View Score Matrix (CVS). Matrix element $DS[m, n]$ is the number of clickstreams in C in which the web page m followed immediately after the web page n . The Matrix element $CVS[m, n]$ is the number of occurrences in which the web page m and the web page n belong to the same clickstream of C . Therefore, DS is not a symmetric matrix, whereas CVS is. The matrix DS represents an adjacency matrix of a weighted directed graph (V, E_{DS}) whose set of vertices V represent web pages or items in the particular system and E_{DS} represents weighted edges from DS matrix. The matrix CVS represents an adjacency matrix of a weighted undirected graph (V, E_{CVS}) , whose set of vertices V represent web pages or items in the particular system and E_{CVS} represents weighted edges from CVS matrix.

Unbiased random walk

In this subsection, we consider a case of a random walk on the large connected component in the non-weighted and undirected graph. We transform the weighted CVS graph to non-weighted graph. Random walk starts at an arbitrary vertex v_k on CVS graph. At each discrete time step n , random walker resides at some vertex v_m and randomly chooses an adjacent vertex v_l from a uniform distribution of adjacent vertices. Let $p_i(n)$ denote a probability that the random walker will be at vertex v_i at discrete time n :

$$p_i(n) = \sum_j \frac{CVS_{ij}}{k_j} p_j(n-1)$$

in which k_j denotes the degree of a vertex v_j . In matrix form we can write $\mathbf{p}(n) = CVS \times D^{-1} \mathbf{p}(n-1) = T \mathbf{p}(n-1)$, in which D is a diagonal matrix with the degrees of vertices k_i down the diagonal and T is a transition matrix. Stationary distributions $\mathbf{p}(\infty)$ or just \mathbf{p} can be expressed as:

$$\mathbf{p} = CVS \times D^{-1} \mathbf{p} \implies$$

$$(I - CVS \times D^{-1}) \mathbf{p} = (D - CVS) \times D^{-1} \mathbf{p} = L D^{-1} \mathbf{p} = 0$$

in which matrix L is the Laplacian matrix of the graph CVS and $D^{-1} \mathbf{p}$ is the eigenvector with the corresponding zero eigenvalue. Thus, we know that $D^{-1} \mathbf{p} = c \mathbf{1}$ for every connected graph, in which c is a constant and $\mathbf{1}$ denotes vector whose components are all ones. This implies that the stationary probability distributions are:

$$p_i = ck_i \implies p_i = \frac{k_i}{\sum_j k_j}.$$

Therefore, we conclude that the stationary probability of a random walker being at a vertex v_i is proportional to its degree k_i [15].

Biased random walk

In this subsection we consider a case of a biased random walk [16] on a large connected component in the weighted CVS graph, which is undirected. Elements of the weighted adjacency matrix CVS of are used as biases for random walks. Let us denote $p_i(n)$ probability that the random walker will be at vertex v_i at discrete time n . This probability can be

expressed as:

$$p_i(n) = \sum_j \frac{CVS_{ij}}{\sum_k CVS_{kj}} p_j(n-1).$$

In matrix form we can write $\mathbf{p}(n) = CVS \times Z^{-1} \mathbf{p}(n-1) = \mathbf{p}(n) = T \mathbf{p}(n-1)$, in which Z is a diagonal matrix with the elements $\sum_k CVS_{ki}$ down the diagonal and T is a transition matrix. By analogue with the unbiased case stationary distributions \mathbf{p} can be expressed as:

$$\mathbf{p} = CVS \times Z^{-1} \mathbf{p} \implies p_i = \frac{\sum_k CVS_{ik}}{\sum_m \sum_n CVS_{mn}}.$$

4. MBRW MODEL

The random walk of this model takes place on the DS graph and uses a combination of biases from the DS and CVS matrices. This discrete time Markov chain model has a finite memory of m past states. Initial vertex can be chosen either by stochastic or deterministic rule. Let us denote the initial vertex as u_1 , then the model chooses the next adjacent vertex u_2 with the probability of:

$$P_{\{u_2|u_1\}} = \frac{DS_{u_2, u_1}}{\sum_k DS_{k, u_1}}$$

thus generating a clickstream $c^i = \{u_1, u_2\}$. The third lecture, u_3 in the clickstream is chosen with the probability of:

$$P_{\{u_3|u_2, u_1\}} = \frac{DS_{u_3, u_2} CVS_{u_3, u_1}}{\sum_k DS_{k, u_2} CVS_{k, u_1}}$$

thus generating a clickstream $c^i = \{u_1, u_2, u_3\}$. Using a finite memory of size m , we choose the vertex u_n with the probability of:

$$P_{\{u_n|u_{n-1}, \dots, u_{n-m+1}\}} = \frac{DS_{u_n, u_{n-1}} \prod_{k=1}^m CVS_{u_n, u_{n-k-1}}}{\sum_j DS_{j, u_{n-1}} \prod_{k=1}^m CVS_{j, u_{n-k-1}}}$$

thus generating a clickstream $c^i = \{u_1, u_2, u_3, \dots, u_n\}$ at the n -th step of the random walk. Clickstream length L is a random variable sampled from a discrete probability distribution like Poisson, negative binomial, geometric, or from the real clickstream length distribution, if available. Using this model, we generated a set of clickstreams $\tilde{C} = \{\tilde{c}^1, \tilde{c}^2, \dots, \tilde{c}^K\}$. In each of the K independent iterations, we determine the clickstream length l and the initial vertex of the random walk. At the end of each iteration i , random walk path $c^i = \{u_1^i, u_2^i, \dots, u_l^i\}$ determines one clickstream appended to the synthetic clickstream set \tilde{C} .

5. MBRW MODEL WITH THE USER-PROFILE SAMPLING PROCEDURE

Let n and q denote total number of users and items in a particular system, respectively. The clickstream set of such system is denoted with $C = \{c^1, c^2, \dots, c^n\}$, in which c^i represents a clickstream of a particular user i . We can construct a user-item usage matrix A from the clickstream set C by breaking the ordering of items in clickstreams. User-item matrix A is element of $R^{n \times q}$ space and contains non-zeros element $A(i, j)$ if and only if the clickstream c^i of the user i contains the item element j . The user-item matrix can be represented as a bipartite graph B , in which first type of vertices are users and second type of vertices are items.

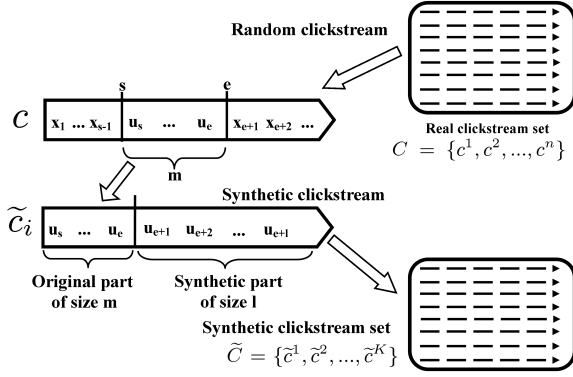


Figure 1: Diagram of MBRW user-profile sampling procedure

Matrix $A^T A$ is a one-mode projection of bipartite graphs to items and is closely related to CVS matrix. Therefore, both DS and CVS matrices represent item preference models and random walks on direct sequence and common view score graphs can only preserve item preferences.

In order to preserve user preferences with random walk models we introduce user-profile sampling procedure as an input memory parameter to the MBRW model. For each synthetic clickstream \tilde{c}^i we want to create, we first randomly sample one real clickstream c^j from the clickstream set C . Let us assume that the sampled real clickstream c^j consists of the following items $\{x_1, \dots, x_{s-1}, u_s, \dots, u_e, x_{e+1}, \dots\}$. Then we create a profile $p^j = \{u_s, \dots, u_e\}$ of user j by taking some random sequence subsample of clickstream c^j between start index s and end index e , which are random variables. In Figure 1 we can see diagram of user-profile sampling procedure. In the next section, we will explain how to sample indexes s and e in more details. Then, we use created user-profile sequence $p^j = \{u_s, \dots, u_e\}$ as a memory component to MBRW model. More precisely, we take the user-profile p^j as the first part of synthetic clickstream $\tilde{c}^i = \{u_s, \dots, u_e\}$ and choose the next item u_{e+1} with the probability of:

$$P_{\{u_{e+1}|u_e, u_{e-1}, \dots, u_s\}} = \frac{DS_{u_{e+1}, u_e} \prod_{k=1}^m CVS_{u_{e+1}, u_{e-k-1}}}{\sum_{z \in Z} DS_{z, u_e} \prod_{k=1}^m CVS_{z, u_{e-k-1}}},$$

in which Z denotes the set of neighbours of item u_e in DS graph. If Z set is empty set, we choose the next item u_{e+1} with the probability of:

$$P_{\{u_{e+1}|u_e, u_{e-1}, \dots, u_s\}} = \frac{CVS_{u_{e+1}, u_e} \prod_{k=1}^m CVS_{u_{e+1}, u_{e-k-1}}}{\sum_{z \in Z} CVS_{z, u_e} \prod_{k=1}^m CVS_{z, u_{e-k-1}}},$$

in which Z denotes the set of neighbours of item u_e in CVS graph. We then continue to use MBRW model until we have generated l new synthetic items. Note, that the value l is also a random variable which is sampled from some probability distribution L . In the end we have generated a synthetic clickstream $\tilde{c}^i = \{u_s, \dots, u_e, u_{e+1}, \dots, u_{e+L}\}$ with first $e - s$ original items and l new synthetic items. As a result of using user-profile sampling with MBRW model we generate synthetic clickstream set \tilde{C} satisfying the user and item preferences. The pseudo code for clickstream MBRW algorithm with the user-profile sampling procedure is provided

in Algorithm 1.

Algorithm 1 Clickstream MBRW model with the user-profile sampling procedure

Input: $C = \{c^1, c^2, \dots, c^n\}$ - real clickstream set,
 DS - Direct Sequence Matrix,
 CVS - Common View Score matrix,
 K - number of synthetic clickstreams,
 Θ - anonymization constant,
 E, M, L : probability distributions
Output: $\tilde{C} = \{\tilde{c}^1, \tilde{c}^2, \dots, \tilde{c}^K\}$ synthetic clickstream set
 $\tilde{C} = \emptyset$

```

for  $i = 1 : K$  do
   $c$  = take random clickstream from  $C$ ;
  //  $c = \{x_1, x_2, \dots, x_{s-1}, u_s, \dots, u_e, x_{e+1}, x_{e+2}, \dots\}$ 
   $e$  = sample "end index" from p.d.  $E$ ;
   $m$  = sample "size of memory" from p.d.  $M$ ;
   $s = e - m$ ; // generate start index
   $profile$  = subsample of  $c$  from  $s$  until  $e$ ;
   $\tilde{c}_i = profile$ ;
  //  $\tilde{c}_i = \{u_s, \dots, u_e\}$  - temporary clickstream
   $l$  = sample number of hops from p.d.  $L$ ;
  for  $j = 1 : l$  do
     $Z$  = find neighbours of item  $u_{e+j-1}$  in  $DS$  graph;
    if  $Z$  not empty set then
      Choose the next item  $u_{e+j}$  with the probability of:
      
$$P_{\{u_{e+j}|\tilde{c}_i\}} = \frac{DS_{u_{e+j}, u_{e+j-1}} \prod_{k=1}^m CVS_{u_{e+j}, u_{e-k-1}}}{\sum_{z \in Z} DS_{z, u_{e+j-1}} \prod_{k=1}^m CVS_{z, u_{e-k-1}}}$$

    else
       $Z$  = find neighbours of item  $u_{e+j-1}$  in  $CSV$  graph;
      Choose the next item  $u_{e+j}$  with the probability of:
      
$$P_{\{u_{e+j}|\tilde{c}_i\}} = \frac{CVS_{u_{e+j}, u_{e+j-1}} \prod_{k=1}^m CVS_{u_{e+j}, u_{e-k-1}}}{\sum_{z \in Z} CVS_{z, u_{e+j-1}} \prod_{k=1}^m CVS_{z, u_{e-k-1}}}$$

    end if
    if  $Z$  empty set then
      end clickstream  $\tilde{c}_i$  and break;
    end if
     $\tilde{c}_i = \tilde{c}_i \cup u_{e+j}$ ; // append new item
  end for
  //  $\tilde{c}_i = \{u_s, \dots, u_e, u_{e+1}, \dots, u_{e+l}\}$ 
  if  $\exists c \in C : sim(\tilde{c}_i, c) \geq \Theta$  then
    // de-anonymization criterion
    discard clickstream  $\tilde{c}_i$  and continue;
  else
     $\tilde{C} = \tilde{C} \cup \tilde{c}_i$ ; - append new synthetic clickstream
  end if
end for

```

6. EXPERIMENTS AND ANALYSIS

Here, we state five hypotheses about the properties of our clickstream MBRW model with the user-profile sampling procedure:

1. Basic statistical properties of \widetilde{DS} and \widetilde{CVS} matrices of synthetic clickstream set \tilde{C} are preserved if we generate a synthetic clickstream set of sufficiently large size.

2. Memory property of our model increases the probability of choosing the relevant next item.
3. User-preferences are largely preserved due to the user-profile sampling procedure although random walk takes place on item-preference graphs even for very small values of memory parameter.
4. Synthetic clickstream sets can be used by Recommender Systems to learn a model on an anonymized dataset while achieving high recommendation performance on real test data.
5. We can provide guarantee that there exist no synthetic clickstream that has similarity to some real clickstream above some threshold Θ .

Now, we describe all the experiments we made in order to confirm our hypotheses. In these experiments we used a sample dataset created from the Yahoo! Music community's preferences to various musical items, released for the KDD-Cup 2011 [43]. We downsampled the dataset, and created a sub-sample of the modest dimensions (10000 users over 5000 items), in order to reduce computational load in numerous experiments. The set of items rated by users ordered in ascending time order represents a set of original clickstreams in our experiments. We use "vertical" and "horizontal" splits of the clickstream dataset to create training and test datasets. Diverse splits were necessary to prove different hypotheses. "Horizontal split" of clickstream dataset C creates two disjoint clickstreams sets C_{train} and C_{test} of fixed sizes. "Vertical split" of clickstream dataset C creates two clickstream sets by a temporal cut. More precisely all the items in a clickstream prior to some specific time t are put to the C_{train} set and all the items in a clickstream after time t are put in C_{test} set. We also apply both splits together to create two disjoint sets of clickstreams of fixed sizes and then make a vertical split on one of them again. We explain for each experiment, the logic of the particular split. These splits are graphically represented in the Figure 2.

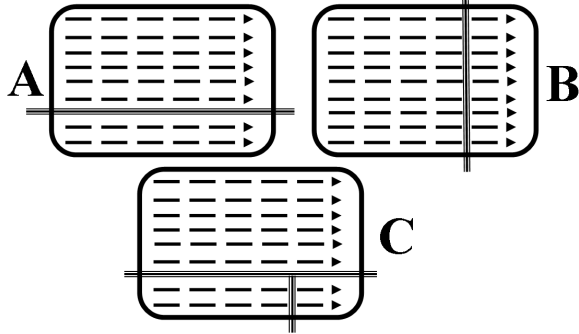


Figure 2: Three ways of splitting the original clickstream set used in computational experiments: A - Horizontal split, B - Vertical split and C - Horizontal and vertical split

Matching basic statistical properties experiment

In this experiment we examined how statistical properties of the item preference matrix like DS and CVS are preserved

in synthetic clickstream set with respect to the original clickstream set. We used "horizontal split" of our clickstream dataset C to two sets C_{train} and C_{test} of 9000 and 1000 clickstreams sizes, respectively. Then, we calculated the DS and CVS matrices of C_{train} dataset and created the synthetic clickstream set \tilde{C} by using MBRW model with user-profile sampling procedure. We used end index parameter e which was sampled from uniform distribution on range from 1 to length of particular sampled clickstream, memory parameter m which was sampled from Gaussian distribution $G(3, 2)$, number of random walk hops parameter l which was sampled from Gaussian distribution $G(9, 2)$ and number of synthetic clickstreams parameter K varying from $10^4 - 10^6$. When we have synthetic clickstream set \tilde{C} , we calculate its statistical properties \widetilde{DS} and \widetilde{CVS} and compare it to the statistical properties of original dataset DS and CVS . It turns out that the various matrix norms like Frobenius and max norm on the difference matrices $(DS - \widetilde{DS})$ and $(CVS - \widetilde{CVS})$ are not appropriate as they do not capture order or ranking preservation between corresponding clickstreams. Therefore we used Spearman's rank correlation measure between the corresponding rows in (DS, \widetilde{DS}) and (CVS, \widetilde{CVS}) .

Table 1: Average rank correlation between (DS, \widetilde{DS}) and (CVS, \widetilde{CVS}) for different sizes (K) of generated synthetic clickstream set. Synthetic clickstream set is created using parameter m sampled from Gaussian distribution $G(3, 2)$, parameter l sampled from Gaussian distribution $G(9, 2)$ and parameter e sampled from Uniform distribution on clickstream length.

Size	$AVG[r(DS, \widetilde{DS})]$	$STD[r(DS, \widetilde{DS})]$
$K = 10^4$	0.5700	0.3210
$K = 5 * 10^4$	0.8261	0.3060
$K = 10^5$	0.8914	0.2224
$K = 5 * 10^5$	0.9308	0.0639
$K = 10^6$	0.9294	0.0590
	$AVG[r(CVS, \widetilde{CVS})]$	$STD[r(CVS, \widetilde{CVS})]$
$K = 10^4$	0.4545	0.2677
$K = 5 * 10^4$	0.5530	0.2407
$K = 10^5$	0.6050	0.2120
$K = 5 * 10^5$	0.7071	0.1765
$K = 10^6$	0.7361	0.1784

Due to the fact that these matrices are sparse and that in the process of recommendation only top ranked items are relevant, we have used rank correlation only for the first $z = 100$ important elements. Rank correlation between complete rows would be misleadingly high due to the row sparsity. Average rank correlation coefficient $AVG[r(DS, \widetilde{DS})] = 0.92$ and $AVG[r(CVS, \widetilde{CVS})] = 0.73$ over all corresponding rows was obtained for first z most important elements, with the above parameters and $K = 10^6$. The rank correlation coefficients for different values of parameter K can be seen in Table 1. This shows highly correlated statistical properties (DS, \widetilde{DS}) and (CVS, \widetilde{CVS}) . Although intuitively clear, it is still important to stress that high rank correlation cannot

be achieved using arbitrary DS and CVS matrices. When two the matrices DS and CVS are constructed from the same clickstream set C , we say that they are "aligned". In order to achieve high similarity between both statistics (DS , \widetilde{DS}) and (CVS , \widetilde{CVS}), the biases for random walk DS and CVS need to be "aligned". We have generated synthetic clickstream set \widetilde{C}_{na} with the $K = 10^5$ with two matrices DS_{na} and CVS_{na} which are not "aligned" (not constructed from the same clickstream set C). In this case we got low similarity between ($CVS_{na}, \widetilde{CVS}_{na}$) and high similarity between ($DS_{na}, \widetilde{DS}_{na}$) (see Table 2).

Table 2: Statistical properties of ($CVS_{na}, \widetilde{CVS}_{na}$) and ($DS_{na}, \widetilde{DS}_{na}$) for "non-aligned" matrices and $K = 10^5$, parameter m sampled from Gaussian distribution $G(3, 2)$, parameter l sampled from Gaussian distribution $G(9, 2)$, parameter e sampled from Uniform distribution of clickstream length

	<i>AVGr</i>	<i>STDr</i>
$(CVS_{na}, \widetilde{CVS}_{na})$	0.1844	0.1706
$(DS_{na}, \widetilde{DS}_{na})$	0.8921	0.2048

Importance of memory - vertical split experiments

In this experiment we did a "vertical split" of our real clickstream set C by a cut t in the sequence of items and created two datasets: C_{train} and C_{test} , prior and after the time t , respectively. Through this experiment we determine how different values of memory affect the probability to generate the relevant item from C_{test} set. More precisely, we create DS and CVS matrices from whole clickstream set C and for each clickstream c_i from C_{train} set with the memory m we calculate the probability to choose the first next relevant item in C_{test} set. For example, if we have clickstream $c_{train}^i = \{u_1, u_2, u_3, u_4, u_5, u_6\}$ from C_{train} and the corresponding clickstream $c_{test}^i = \{u_7, u_8, \dots\}$ in C_{test} then the first next relevant item for user i is u_7 . Relevant next item probability is defined as the probability of choosing the item u_7 over all other items in MBRW model by using memory m . As expected, we observe that by introducing the memory to random walk model we can increase the average relevant item probability by approximately 60 percent on this dataset (see Figure 3).

To assess more realistic situation, that of a real recommender system, in which we assume only partial knowledge of C dataset, thus also a partial knowledge of DS and CVS matrices we perform another "vertical split" experiment. In this experiment we do the same temporal split like in the previous experiment and create two datasets: C_{train} and C_{test} . Here, we create DS and CVS from C_{train} dataset only, in contrast to the previous experiment in which we created them from whole C dataset. This means that we got the partial information and still measure the relevant next item probability. Note, that in this case of partial DS and CVS matrices the Z set of neighbouring nodes of last item in clickstreams of C_{train} in DS graph can be empty set. Then we calculate the relevant item probability by walking in CVS graph. From the Figure 3 we can see that by intro-

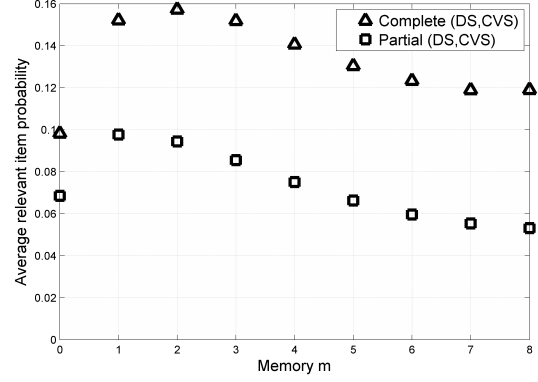


Figure 3: Average relevant item probability with respect to memory of MBRW model, when complete and partial information about the DS and CVS matrices are available

ducing the memory to random walk model, we can increase the average relevant item probability by approximately 42 percent on this dataset by using partial information of DS and CVS matrices.

Assessing preference profile preservation

In this experiment we want to see how well can MBRW model with user-profile sampling procedure as input reproduce user preference structure of the original clickstream set. We define profile of user i to be any subsequence of clickstream c_i in the clickstream set C . We created a user-item usage matrix A from the real clickstream set C . Remember, that matrix A contains non-zero elements $A(i, j)$ if and only if user i consumed item j . Typically, matrix A is very sparse, thus in order to compare more efficiently original and synthetic usage matrix, or virtual-user profiles, we mitigated the sparsity problem by projecting into a dense reduced dimensional space using matrix factorization. Singular Value Decomposition (SVD) is a matrix factorization technique of an $m \times n$ matrix A of the form: $A = USV^T$. If A is a real matrix, SVD is a factorization of the form $A = USV^T$, in which U is a unitary matrix of left singular vectors of A , matrix V is a unitary matrix of right singular vectors of A , and matrix S is a diagonal matrix whose non-zero elements are singular values of A . Best r low rank approximation with respect to the Frobenius norm of matrix A is matrix which takes r greatest singular values and left and right singular vectors to obtain matrix $A_r = U_r \times S_r \times V_r^T$. The last statement is known as the Eckart-Young theorem [17]. The matrix U_r represents user-preference in r dimensional latent space and the matrix V_r represent items in r dimensional latent space [18].

To ensure mapping between the real and the synthetic user, we set the number of synthetic and real users to be equal ($K = n$), and changed the sampling procedure to sample each real clickstream only once. Then, we created synthetic clickstream set \widetilde{C} and corresponding user-item usage matrix \widetilde{A} . Afterwards, we permuted the rows of matrix \widetilde{A} so that corresponding rows represent same users as matrix A . Note, when the numbers of artificial and real users are

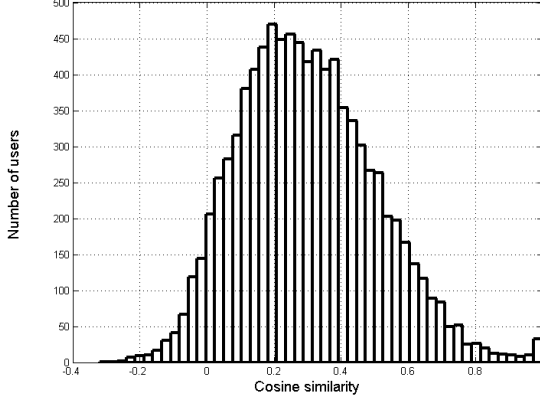


Figure 4: Histogram of cosine similarity between corresponding rows of real user-preferences U_k and synthetic user-preference \tilde{U}_k in reduced latent feature space $r = 100$. Synthetic clickstream was generated with parameter m sampled from Gaussian distribution $G(3, 2)$ and parameter l sampled from Gaussian distribution $G(5, 2)$ and $K \approx 9000$.

equal, the constraint that each real clickstream can be sampled only once is a good approximation because user-profile sampling procedure uses Uniform distribution.

To compare real and synthetic user-preferences we do a fold-in mapping [19] of every user vector u_i from \tilde{A} to r dimensional latent feature space of matrix A_r . Mapping synthetic user vectors \tilde{A}_r to \tilde{U}_r is made by the following transformation:

$$\tilde{U}_r = \tilde{A}_r * V_r * S_r^{-1}.$$

Now, we have two matrices (U_r, \tilde{U}_r) of dimension $n \times r$ in latent low dimensional space representing real and synthetic user-preference. We use cosine similarity on vectors between the corresponding rows of U_r and \tilde{U}_r matrices.

In Figure 4 we present histogram of cosine similarity between corresponding rows of real user-preferences U_r and synthetic user-preference \tilde{U}_r in reduced latent feature space. Cosine similarity ranges from -1 indicating dissimilarity to 1 indicating similarity. Figure 4 shows that there is significant positive offset in similarity which reflects preservation of user preferences.

To gain more insight, we experimented by changing parameters m and l over a set of deterministic values. For that reason we have generated 121 different synthetic clickstreams sets \tilde{C}_{ml} with parameters m and l varying in interval $[0-10]$. Figure 5 depicts that only small amount of memory is needed so that the average cosine similarity between the real and the synthetic users becomes significantly higher. Figure 6 depicts that average cosine similarity between the real and the synthetic users drops very slowly when we increase number of random hops.

Recommender system performance experiment

In this set of experiments we tested the quality of synthetically generated clickstreams by comparing performance of

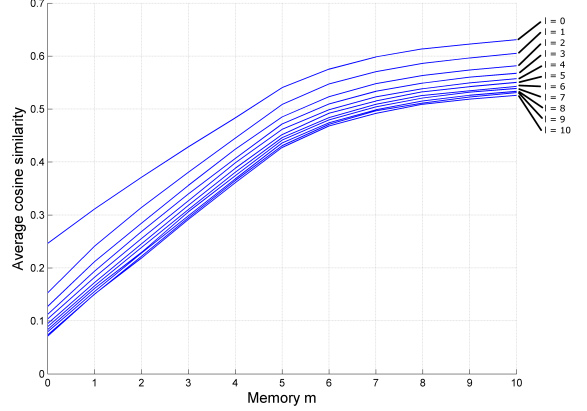


Figure 5: Average cosine similarity in reduced latent feature space $r = 100$ over all corresponding pairs of real and synthetic users and $K \approx 9000$. Memory parameter m and number of random hops parameter l are varying in interval $[0-10]$.

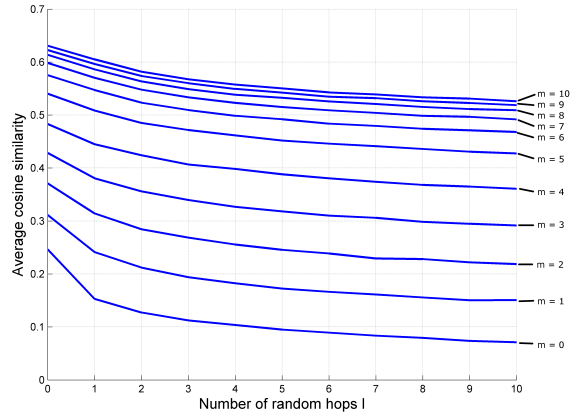


Figure 6: Average cosine similarity in reduced latent feature space $r = 100$ over all corresponding pairs of real and synthetic users and $K \approx 9000$. Memory parameter m and number of random hops parameter l are varying in interval $[0-10]$.

two recommendation models based on original C and synthetic set of clickstreams \tilde{C} . Figure 7 illustrates the experimental setup. We have used "horizontal split" of our clickstream dataset C to two sets C_{train} and C_{test} of 9000 and 1000 clickstreams sizes, respectively. We use Item-based k-nn algorithm [20] as recommender to generate models for the comparison. This algorithm takes user-item usage matrix A_{train} , constructed from C_{train} dataset and generates a model M . Model M is a item to item similarity matrix reduced in a way that each column i contains only k most similar non-zero elements. Then, we take C_{test} clickstream set and do an additional "vertical or temporal split" to two sets C_{query} and C_{result} (see Figure 2, C part). Item-based k-nn algorithm produces top-N recommendation list for each clickstream in C_{query} dataset and evaluates predictions against ground truth solution C_{result} using $Recall@n$, measure used often in information retrieval. $Recall@n$ tell us what is the probability that a relevant items is retrieved in top-N recommendation list.

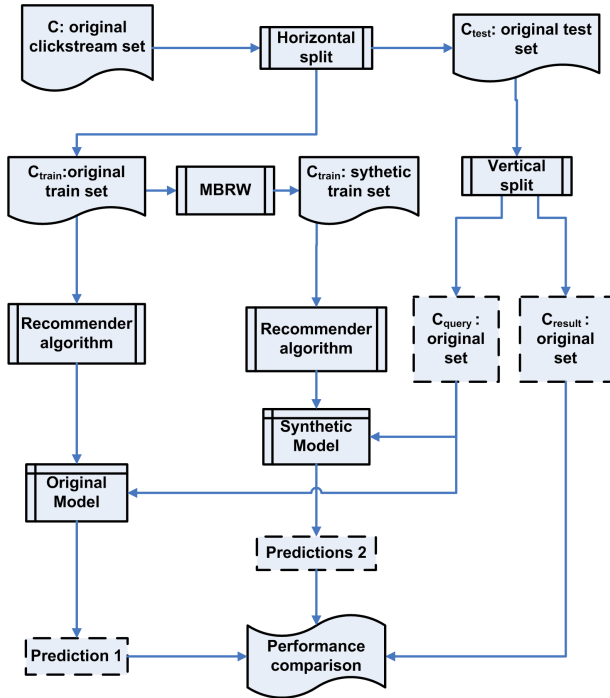


Figure 7: Experimental diagram for testing the recommender system performance on original C and synthetic set of clickstreams \tilde{C} .

In first experiment of this kind we generated the synthetic clickstream set \tilde{C}_{train} from C_{train} with parameters m sampled from Gaussian distribution $G(3, 2)$ and l sampled from Gaussian distribution $G(5, 2)$ and $K \approx 9000$. Item-based k-nn algorithm ($k=15$) is then used to build two recommender models M and \tilde{M} using C_{train} and \tilde{C}_{train} , respectively. These models are then used to generate predictions for the same query set C_{query} , and obtained results are evaluated against C_{result} using $Recall@5$ evaluation measure. Figure 8 gives the results of these two models represented as frequency histograms over particular $Recall@5$ values. High agreement in histogram profile implies that original user

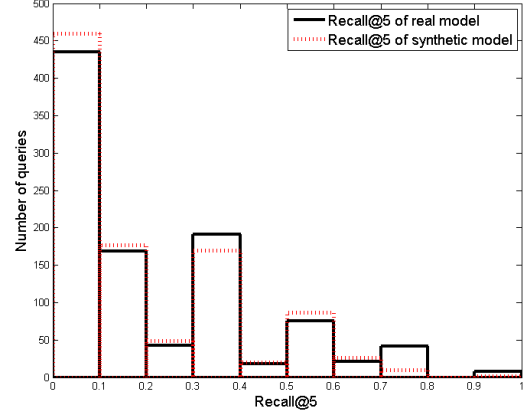


Figure 8: Histograms of $recall@5$ performance of RS with real model M and synthetic model \tilde{M} . Average $recall@5$ for $RS(M)=0.2251$ and for $RS(\tilde{M})=0.2021$.

profiles are well preserved in synthetic clickstream dataset \tilde{C}_{train} .

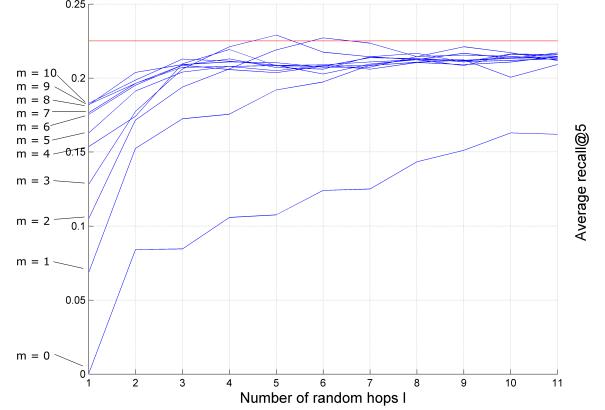


Figure 9: Average $recall@5$ for different models \tilde{M}_{ml} on C_{query} and C_{result} clickstream datasets and performance of real model M (red curve)

To get more detailed picture of the performance of recommender systems, for synthetic datasets generated with different deterministic values of parameters m and l , we have generated 121 different synthetic clickstreams sets \tilde{C}_{ml} with parameters m and l varying in interval $[0, 10]$. For each synthetic clickstream set \tilde{C}_{ml} we made a different model of \tilde{M}_{ml} and evaluate its performance as in the previous experiment. In Figure 9 we can see the average $recall@5$ performance of recommender system that uses \tilde{M}_{ml} on the C_{query} and C_{result} clickstream dataset. Relatively steep rise in performance of the synthetically generated models with $m > 3$ and $l > 3$ show that it is possible to generate recommendation models similar to those generated using original clickstreams.

User-based k-nn	AUC	prec@5	prec@10	prec@15	NDCG	MAP
Real performance	0,93	0,20	0,17	0,15	0,47	0,19
Synthetic performance	0,87	0,17	0,15	0,13	0,43	0,15
Item-based k-nn	AUC	prec@5	prec@10	prec@15	NDCG	MAP
Real performance	0,93	0,26	0,23	0,21	0,52	0,26
Synthetic performance	0,87	0,21	0,18	0,17	0,47	0,20
BMF-Factorization	AUC	prec@5	prec@10	prec@15	NDCG	MAP
Real performance	0,95	0,21	0,15	0,12	0,46	0,19
Synthetic performance	0,92	0,21	0,16	0,13	0,46	0,19
WRM-Factorization	AUC	prec@5	prec@10	prec@15	NDCG	MAP
Real performance	0,91	0,27	0,23	0,21	0,53	0,26
Synthetic performance	0,83	0,19	0,17	0,15	0,46	0,18

Table 3: Performances of different recommender systems using models generated from original and synthetic clickstreams, tested against same test set C_{result} .

Final set of recommender quality experiments involved assessment of synthetically generated recommendation models for diverse algorithms [22] [21] (Weighted user-based k-nn [26], Weighted item-based k-nn [20], Biased Matrix Factorization [23], Weighted Regularized Matrix Factorization [25]) and using other more different measures (AUC, NDCG, MAP, precision). The results are given in the Table 3 for the synthetic clickstreams generated using MBRW with parameters m sampled from Gaussian distribution $G(3, 2)$, l sampled from Gaussian distribution $G(5, 2)$ and $K=9000$.

7. ANONYMIZATION

Most of the anonymization strategies fail to guarantee privacy on real datasets. Real world datasets like recommendations, preferences, transactions, etc. tend to be high dimensional and mostly sparse. Furthermore, de-anonymization attack algorithms can use any available background knowledge along with probabilistic reasoning to breach privacy. Narayanan and Shmatikov [7] shown that there exist limits of privacy in public datasets. They, also constructed a formal model for privacy breaches in anonymized datasets. We provide only basic notion and definitions from their work [7] to demonstrate anonymization capabilities of the MBRW model. We start with the definition of (ϵ, θ) -sparsity of the database. A database D is (ϵ, θ) -sparse w.r.t. the similarity measure sim if:

$$P(sim(r, r') > \epsilon, \forall r' \neq r) \leq \theta.$$

Similarity measure is a function that maps two user vectors in dataset to the interval $[0, 1]$.

A database D can be (Θ, ω) de-anonymized w.r.t. auxiliary information Aux if there exists an algorithm A which, on inputs D and $Aux(r)$ in which $r \in D$ outputs r' such that:

$$P(sim(r, r') \geq \Theta) \geq \omega.$$

Our MBRW model with user-sampling procedure generates partially synthetic dataset. Due to the fact that these clickstreams are results of discrete stochastic process privacy is preserved in a way that re-identification is uncertain. However, MBRW approach also provides re-identification protection guarantee w.r.t. (Θ, ω) de-anonymization definition via simple clickstream privacy filtering.

To illustrate the procedure of clickstream privacy filtering we have generated the synthetic clickstream set \widetilde{C}_{train} from C_{train} with parameters m sampled from Gaussian dis-

tribution $G(3, 2)$ and parameter l sampled from Gaussian distribution $G(5, 2)$ and $K \approx 25000$. We have generated $K \approx 25000$ synthetic clickstreams from the original clickstream set C_{train} , which contains ≈ 9000 real clickstreams. We can then simulate de-anonymization attack by assuming that attacker possess perfect knowledge of the original dataset C_{train} , used to generate our synthetic dataset. In order to provide de-anonymization protection guarantee for our synthetic dataset it is sufficient to calculate the similarity between all pairs of clickstreams from \widetilde{C}_{train} and C_{train} , and filter out those clickstreams from \widetilde{C}_{train} that have similarity over some threshold Θ . To illustrate the proportions of the dataset that has similarity to original clickstreams above some threshold Θ , we have calculated all the pairwise similarities between the original and synthetic clickstreams. The distribution is depicted in Figure 10. For similarity function we have used cosine similarity between vectors of users. After pruning the clickstream \widetilde{C}_{train} set we guarantee that there exist no synthetic clickstream that has similarity to some real clickstream above some threshold Θ or that :

$$P(sim(r, r') \geq \Theta) = 0.$$

If for example we set the threshold Θ to be 0.7, then after pruning, the synthetic clickstream \widetilde{C}_{train} set will contain approximately 80% of its original size $K \approx 25000$. Note, that we have assumed that the attacker obtained 100% of real dataset and that the similarity between clickstreams is calculated as similarity of binary vectors where ordering is not relevant. In real life scenario attacker typically has access to small sample of original clickstream set along with their identification attributes. In our experiments we obtained small impact on recommender performance for Θ as low as 0.60.

8. CONCLUSIONS

The principle aim of our work was to construct generator of real-like clickstream datasets, that preserve structure of original user-item preferences but is at the same time addressing privacy protection requirements (resilience to "linkage" attacks). With respect to this aim we have investigated properties of the memory biased random walk model with the user-profile sampling procedure.

Infinite many sequences or paths can be constructed from clickstream graph, but with memory biased random walk model (MBRW) we sample sequences that are more likely to

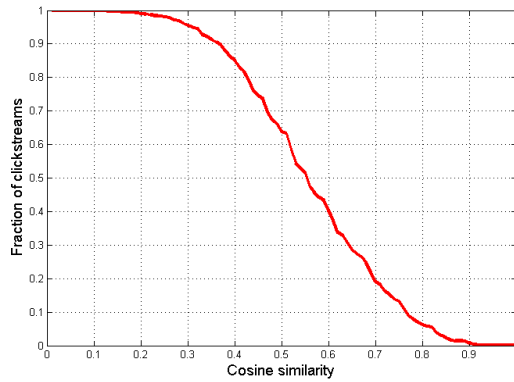


Figure 10: Fraction of synthetic clickstreams having similarity to most similar original clickstream above some value, for the simulated dataset.

occur in a system described by DS , CVS statistics. Furthermore, we sample clickstreams according to sampled user-profiles in order to preserve user-preferences in a system. We demonstrated that the basic statistical properties of "aligned" data generators DS and CVS matrices are preserved in synthetic dataset if we generate dataset of sufficiently large size. We demonstrated that memory property of MBRW model increases the probability of choosing the relevant next item over all other items in a system.

In addition to presenting the new algorithm for synthetic clickstream generation, we demonstrate that synthetic datasets created with MBRW model can be used to learn recommender models and achieve high recommendation performance on the real data. At the same time this approach is amenable to simple mechanism for re-identification protection in the sense of the (Θ, ω) de-anonymization guarantee.

Important impacts of these results we see for the research, whereas by combining high quality replica of usage data and item content descriptions more realistic datasets could be made available for research. With the same approach commercial enterprises that rely on outsourcing in recommender system development, could benefit through lowering their privacy-breaching risks.

9. ACKNOWLEDGMENTS

This work is supported by the European Community 7th framework ICT-2007.4 (No 231519) "e-LICO: An e-Laboratory for Interdisciplinary Collaborative Research in Data Mining and Data-Intensive Science". V.Z. acknowledges support from Croatian ministry of education, science and sport project No. 098-0352828-2863 and from EU FET Open Project FOC nr 255987.

10. REFERENCES

- [1] RENDLE, S., TSO-SUTTER, K., HUIJSEN, W., FREUDENTHALER, C., GANTNER, Z., WARTENA, C., BRUSSEE, R. AND WIBBELS, M., "Report on State of the Art Recommender Algorithms (Update)", *MyMedia public deliverable D4.1.2.*, March 2011
- [2] ADOMAVICIUS, G., AND TUZHILIN, A., *Towards the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions*, *IEEE Transactions on knowledge and data engineering*, vol. 17, no.6, June 2005.
- [3] SALTON, G., *Automatic Text Processing The Transformation, Analysis, and Retrieval of Information by Computer*, Addison Wesley, 1989
- [4] BAEZA-YATES, R. AND RIBEIRO-NETO, B., *Modern Information Retrieval*, Addison Wesley, May 1999
- [5] RESNICK, P., IAKOVOU, N., SUSHAK, M., BERGSTROM, P., AND RIEDL, J., *GroupLens: An Open Architecture for Collaborative Filtering of Netnews*, *Proc. 1994 Computer Supported Cooperative Work Conf.*, 1994
- [6] BURKE, R., "Hybrid recommender systems: Survey and experiments", *User Modeling and User-Adapted Interaction*, 12(4):331–370, 2002
- [7] NARAYANAN, A., SHMATIKOV, V., *Robust Deanonimization of Large Sparse Datasets (How to Break Anonymity of the Netflix Prize Dataset.)* *IEEE S&P* 2008
- [8] SWEENEY, L., *Achieving k-anonymity privacy protection using generalization and suppression*, *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10 (5), 2002; 571-588
- [9] Gori, M., and Pucci, A., *Research Paper Recommender Systems: A Random-Walk Based Approach*, *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*
- [10] Abbassi, Z. and Mirrokni, V.S., *A recommender system based on local random walks and spectral methods*, *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*
- [11] Bogers, T., *Movie Recommendation using Random Walks over the Contextual Graph*, *Proceedings of the 2nd Workshop on Context-Aware Recommender Systems*, (2010)
- [12] Fouss, F., Saerens, M., Pirotte, A., Kolp, M., and Faulkner, S., *Web recommendation system based on a Markov-chain model*, *International Conference on Enterprise Information Systems (ICEIS 2005)*, Miami, USA)
- [13] FELLER, W., *An Introduction to Probability theory and its applications*, John Wiley and sons, New York London Sydney (1968), ISBN: 9780471257080
- [14] KAO, E. P. C., *An Introduction to Stochastic Processes*, Duxbury Press, (1996), ISBN: 0534255183
- [15] NEWMAN, M. E. J., *Networks An Introduction*, Oxford University Press, (2010), ISBN: 9780199206650
- [16] ZLATIĆ, V., GABRIELLI, A., CALDARELLI, G., *Topologically biased random walk and community finding in networks*, *Phys. Rev. E* 82, 066109 (2010)
- [17] Stewart, G. W., *On the Early History of the Singular Value Decomposition*. *SIAM Review* 35 (4): 551-566, (1993)
- [18] SARWAR, B., KARYPI, G., KONSTA, J. AND RIEDL, J., "Application of Dimensionality Reduction in Recommender System - A Case Study", In *ACM WebKDD Workshop*, 2000
- [19] DEERWESTER, S., DUMAS, S., FURNAS, G., LANDAUER, T., HARSMAN, R., "Indexing by latent

- semantic analysis*”, *Society for Information Science*, 41, pp. 391, 1990
- [20] DESHPANDE, M. AND KARYPIS, G., *Item-based top-n recommendation algorithms*. *ACM Transactions on Information Systems*, 22(1):143-177, 2004
 - [21] MIHELICIC, M., SMUC, T., *Recommender Extension for RapidMiner*, January 2012, http://rapidupdate.de:8180/UpdateServer/faces/product_details.xhtml?productId=rmx_irbrecommender.
 - [22] GANTNER, Z. RENDLE, S., AND FREUDENTHALER, C., *MyMediaLite Recommender System Library*, January 2011.
 - [23] RENDLE, S., FREUDENTHALER, C., GANTNER, Z., SCHMIDT-THIEME, L., *BPR: Bayesian Personalized Ranking from Implicit Feedback*. *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI 2009)*, Montreal, Canada, 2009
 - [24] HU, Y., KOREN, Y., VOLINSKY, C., *Collaborative filtering for implicit feedback datasets*, *IEEE International Conference on Data Mining (ICDM)*, 2008
 - [25] PAN, R., ZHOU, Y., CAO, B., LIU, N. N., LUKOSE, R.M., SCHOLZ, M., YANG, Q., *One-class collaborative filtering*, *IEEE International Conference on Data Mining (ICDM)*, 2008
 - [26] SARWAR, B. M., KARYPIS, G., KONSTAN, J.A., RIEDL, J., *Analysis of recommendation algorithms for e-commerce*. In *Proceedings of the Second ACM Conference on Electronic Commerce (EC’00)*, pages 285-295, 200
 - [27] FUNG, B.C.M., WANG, K., WAI-CHEE FU, A., YU, P. S., *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*, Chapman & Hall/CRC, 2011, ISBN-10: 1420091484
 - [28] CHEN, B.-C., KIFER, D., LEFEVRE, K., *Privacy-Preserving Data Publishing*, *Foundations and Trends In Databases* 2:1-2(2009), ISBN-10: 1601982763
 - [29] AGGARWAL, C. C., YU, P.S., *Privacy-Preserving Data Mining: Models and Algorithms (Advances in Database Systems)*, Springer Science+Business Media LLC, 2008, ISBN-10: 1441943714
 - [30] SAMARATI, P., *Protecting respondents identities in microdata release*, *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 13(6):1010-1027, 2001
 - [31] SAMARATI, P., SWEENEY, L., *Generalizing data to provide anonymity when disclosing information*, In *Proc. of the 17th ACM SIGACTSIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, page 188, Seattle, WA, June 1998
 - [32] SAMARATI, P., SWEENEY, L., *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression*. *Technical report*, SRI International, March 1998
 - [33] MACHANAVAJJHALA, A., GEHRKE, J., KIFER, D., VENKITASUBRAMANIAM, M., *Idiversity: Privacy beyond k-anonymity*. In *Proc. of the 22nd IEEE International Conference on Data Engineering (ICDE)*, Atlanta, GA, 2006.
 - [34] MACHANAVAJJHALA, A., KIFER, D. GEHRKE, J., VENKITASUBRAMANIAM, M., *Idiversity: Privacy beyond k-anonymity*. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1), March 2007.
 - [35] DWORK., C., *Differential privacy*. In *Proc. of the 33rd International Colloquium on Automata, Languages and Programming (ICALP)*, pages 1-12, Venice, Italy, July 2006
 - [36] WOLF, P.P., GOUWEELEEUW, J.M., KOOIMAN, P., WILLENBORG, L., *Reflections on PRAM. Statistical data protection, proceedings of the conference*, Lisbon, 1998
 - [37] WOLF, P.P., GELDER, I., *An empirical evaluation of PRAM*, *Discussion paper 04012, Statistics Netherlands*, September 2004
 - [38] RAGHUNATHAN, T.E., REITER, J.P., RUBIN, D.B., *Multiple Imputation for Statistical Disclosure Limitation*. *Journal of Official Statistics*, Vol. 19, No. 1, 2003, pp. 1-16
 - [39] FIENBERG, S.E., *A radical proposal for the provision of micro-data samples and the preservation of confidentiality*, *Technical report 611*, Carnegie Mellon University Department of Statistics, 1994
 - [40] DANDEKAR, R. A., COHEN, M., KIRKENDALL, N., *Sensitive Micro Data Protection Using Latin Hypercube Sampling Technique*. In *Inference Control in Statistical Databases, From Theory to Practice*, Josep Domingo-Ferrer (Ed.). Springer-Verlag, London, UK, UK, 117-125
 - [41] REITER, J.P., *Inference for partially synthetic, public use microdata sets*, *Survey Methodology*, 29:181-188, 2003
 - [42] DANDEKAR, R.A., DOMINGO-FERRER, J., SEBE, F., *LHS-Based Hybrid Microdata vs Rank Swapping and Microaggregation for Numeric Microdata Protection*. In *Inference Control in Statistical Databases, From Theory to Practice*, Josep Domingo-Ferrer (Ed.). Springer-Verlag, London, UK, UK, 153-162
 - [43] DROR, G., KOENIGSTEIN, N., KOREN, Y., WEIMER, M., *The Yahoo! Music Dataset and KDD-Cup’11*. *KDD-Cup Workshop*, 2011
 - [44] RecLab prize on Overstock.com, January 2012 <http://overstockreclabprize.com/>
 - [45] VENGROFF, D.E., *RecLab-core*, January 2012, <http://code.richrelevance.com/reclab-core/>